Paraphrase Generation and Evaluation on Colloquial-Style Sentences

Sjöblom, Eetu Ilari

European Language Resources Association (ELRA)
2020-05-01

Sjöblom , E I , Creutz , M & Scherrer , Y 2020 , Paraphrase Generation and Evaluation on Colloquial-Style Sentences . in N Calzolari , F Béchet , P Blache , K Choukri , C Cieri , T Declerck , S Goggi , H Isahara , B Maegaard , J Mariani , H Mazo , A Moreno , J Odijk & S Piperidis (eds) , Proceedings of the 12th Language Resources and Evaluation Conference . European Language Resources Association (ELRA) , Paris , pp. 1814-1822 , Language Resources and Evaluation Conference , 11/05/2020 . < https://www.aclweb.org/anthology/2020.lrec-1.224 >

http://hdl.handle.net/10138/326098

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1814–1822
Marseille, 11–16 May 2020

Paraphrase Generation and Evaluation on Colloquial-Style Sentences

Eetu Sjöblom, Mathias Creutz, Yves Scherrer
Department of Digital Humanities, Faculty of Arts, University of Helsinki, Finland
{eetu.sjoblom,mathias.creutz,yves.scherrer}@helsinki.fi

Abstract

In this paper, we investigate paraphrase generation in the colloquial domain. We use state-of-the-art neural machine translation models trained on the Opusparcus corpus to generate paraphrases in six languages: German, English, Finnish, French, Russian, and Swedish. We perform experiments to understand how data selection and filtering for diverse paraphrase pairs affects the generated paraphrases. We compare two different model architectures, an RNN and a Transformer model, and find that the Transformer does not generally outperform the RNN. We also conduct human evaluation on five of the six languages and compare the results to the automatic evaluation metrics BLEU and the recently proposed BERTScore. The results advance our understanding of the trade-offs between the quality and novelty of generated paraphrases, affected by the data selection method. In addition, our comparison of the evaluation methods shows that while BLEU correlates well with human judgments at the corpus level, BERTScore outperforms BLEU in both corpus and sentence-level evaluation.

Keywords: paraphrase generation, colloquial language, neural machine translation, evaluation metrics

1. Introduction

Paraphrases are a set of sentences or phrases that have the same meaning. The study of paraphrases has both theoretical and practical implications: On the one hand, it is possible to explore semantic representations that go deeper than surface-level features. Two expressions may carry the same meaning, although they may not contain the same words or their syntactic structures may be completely different. These two expressions could have an identical or similar underlying semantic representation or there could be a mapping that transforms one surface form to another.

On the other hand, there are practical applications of paraphrase models. Such models can be useful in information retrieval or data mining for discovering expressions with the intended meaning but with totally different surface realization than the original query (Riezler et al., 2007). In addition, paraphrasing is used in abstractive summarization as part of summarization models (Nayeem et al., 2018), as well as for evaluation (Vadapalli et al., 2017). Paraphrases can also be used for proofing or grammar checking, producing suggested corrections. Similarly, someone perfecting their skills in a second language, or someone looking for alternate, possibly more idiomatic, expressions may benefit from paraphrase models. For instance, to pick one word, to *corroborate*, in a few contexts, we can find the following paraphrase pairs: *"She'll corroborate my story."* → *"She'll back me up."*, *"Can you corroborate that?"* → *"I need proofs."*, *"Will people corroborate your account?"* → *"Is there anybody who can vouch for that?"*

In this paper, we focus on paraphrase generation using neural machine translation methods. In paraphrase generation we are interested in models that take in an arbitrary input sentence and generate an output with the same meaning but different surface form. We apply traditional recurrent encoder-decoder networks with attention (Luong et al., 2015) as well as Transformer based models (Vaswani et al., 2017), which are the state of the art of modern machine translation. Previous work has already addressed paraphrase generation through machine translation trained on monolingual data (Quirk et al., 2004; Hasan et al., 2016; Prakash et al., 2016). Variants integrating variational autoencoders into the models (Gupta et al., 2018) or different learning schemes based on reinforcement learning (Li et al., 2017) have also been proposed. Roy and Grangier (2019) propose a method based on variational autoencoders and unlabeled monolingual data. A closely related approach uses machine translation models to generate paraphrases via backtranslation (Mallinson et al., 2017; Suzuki et al., 2017; Wieting and Gimpel, 2018), where a sentence is first translated into one or more target languages and then back into the source language. In addition to machine translation models, rule-based systems (Meteer and Shaked, 1988) and methods based on lexical substitution (Kauchak and Barzilay, 2006) have previously been used for paraphrase generation.

There are other approaches to finding paraphrases beside paraphrase generation. Pivot methods rely on parallel corpora (Bannard and Callison-Burch, 2005) whereas paraphrase detection and extraction can be based on sentence embeddings (Wieting and Gimpel, 2018; Wieting and Gimpel, 2017; Sjöblom et al., 2018) or semantic matching models (Lan and Xu, 2018). With arbitrary input sentences, these methods are intuitively unappealing because they search for a closest match in the available data, and no amount of data guarantees that a correct paraphrase is found for any given input. For this reason, we adopt the generation approach. Naturally, generation models are still dependent on suitable training data, and generalization outside the training set is a problem that needs to be tackled. However, massive data sets beyond the training data are not needed in this approach.

We are interested in paraphrase generation specifically in less formal, colloquial style language. A less formal domain can be especially useful when paraphrasing is used in a language learning context for suggesting alternatives and more idiomatic expressions to second-language learners. This is in contrast to many of the common paraphrase data sets such as the Microsoft Research Paraphrase Cor-

pus consisting of news text (Dolan et al., 2004; Dolan and Brockett, 2005), PPDB (Ganitkevitch et al., 2013), which covers many formal domains, in addition to some more colloquial data, or Quora Question Pairs (Iyer et al., 2017), which covers a variety of topics but is limited to questions. In addition, most of the work on paraphrase generation has been for English, while we are interested in broadening the work to multiple languages. We focus on the Opusparcus corpus for our experiments (Creutz, 2018). Opusparcus consists of sentential paraphrases in six languages extracted from subtitles of movies and TV shows. The English subset of Opusparcus has been previously used in paraphrase generation (Ampomah et al., 2019; Hämäläinen and Alnajjar, 2019), but to our knowledge, no previous work has used all six languages in the corpus.

We perform systematic evaluation and analysis of paraphrase generation. To assess semantic adequacy of the generated paraphrases, we compute scores from manual annotations, which we compare to BLEU (Papineni et al., 2002) and a recently proposed, so-called BERTScore (Zhang et al., 2020). Furthermore, we quantify the novelty of the phrases using PINC scores (Chen and Dolan, 2011).

Major contributions of the present work are experiments in data selection to understand the trade-off between semantic adequacy and novelty in paraphrase generation, as well as the validation of BERTScore in the colloquial domain using manual evaluation. In contrast to much of the previous work, we also perform experiments in multiple languages.

2. Data

Opusparcus (Creutz, 2018) is a sentential paraphrase corpus consisting of pairs of sentences extracted automatically from the OpenSubtitles corpus (Lison and Tiedemann, 2016). Opusparcus is publicly available[1] and contains training, development and test sets for six European languages: German (de), English (en), Finnish (fi), French (fr), Russian (ru), and Swedish (sv).

Training sets: The Opusparcus training sets consist of millions of sentence pairs. These sentence pairs have not been annotated manually, but a ranking function has been used for sorting, such that the sentence pairs that are most likely to be true paraphrases are in the beginning of the data set and the least likely paraphrase pairs are last. This allows us to pick cleaner and smaller training sets or larger but noisier sets. In the present work, we use a threshold such that 70% of the sentence pairs are estimated to be true paraphrases. This threshold has been shown earlier to perform well in paraphrase detection experiments (Sjöblom et al., 2018).

Test sets: For each language, Opusparcus provides a test set of approximately 1000 sentence pairs that have been verified to be "Good" or "Mostly good" paraphrases by human annotators (Creutz, 2018; Aulamo et al., 2019). None of these sentence *pairs* occur in the training sets. However, around half of the test sentences do occur in the training set, although paired with some other sentence than in the test set. This makes sense in a paraphrase *detection* scenario, where a classifier predicts whether a given (unseen) sentence pair consists of paraphrases or not. In a paraphrase *generation* scenario, however, it is unsatisfactory to evaluate performance on sentences that have been observed during training, because appropriate paraphrases could be produced by memorizing the training set.

We therefore divide the test set into two subsets of approximately the same size: the *seen* test set, which consists of sentence pairs where the source sentence is part of the training set, as well as the *unseen* test set, which consists only of sentences that have not been present during training. Most of the evaluation will focus on the unseen test set, which is the most interesting scenario.

[1] http://urn.fi/urn:nbn:fi:lb-2018021221

3. Paraphrase Generation Models

We use two different neural machine translation architectures in our experiments. Our first model is a standard sequence-to-sequence network with general attention (Luong et al., 2015). The encoder first encodes the input sentence into a sequence of vectors using a recurrent neural network (RNN), and the decoder RNN then selectively pays attention to the encoded vectors to generate the output sentence. We use LSTM units in both the encoder and the decoder, with a 3-layer bidirectional encoder and a 3-layer unidirectional decoder. We train separate encoder and decoder word embeddings with 512 dimensions and use 1024 dimensions in the encoder and decoder layers. The total number of parameters in the model is approximately 110 million. A dropout probability of 0.3 is used between the LSTM layers. The parameters were kept constant for all languages.

Our second model is the Transformer model by Vaswani et al. (2017). The Transformer has been successfully adapted to a wide variety of sequence problems and has specifically achieved state-of-the-art results in machine translation. Instead of recurrent connections present in the RNN model, the Transformer is based purely on self-attention within the encoder and the decoder, as well as attention between the encoder and the decoder. We use 6 layers in both the encoder and the decoder, with hidden state and word embedding dimensionalities of 512, separate word embeddings for encoder and decoder, 8 attention heads, and a feed-forward dimensionality of 2048 within the layers. The total number of parameters in the model is approximately 90 million. A dropout of 0.1 is used between layers. These and the rest of the hyperparameters for the Transformer follow the recommended setup of OpenNMT-py (Klein et al., 2017), which we use for all experiments.

All models are trained for 400k steps or until convergence, with a validation score as the convergence criterion. The data is preprocessed using byte pair encoding (BPE) (Sennrich et al., 2016) with 30k operations to avoid out-of-vocabulary tokens. We use the Adam optimizer (Kingma and Ba, 2014) to train both models with a learning rate of 0.0001. For the RNN model, the learning rate is halved every 40k steps starting after 200k steps, and for the Transformer, the recommended noam decay is used. We use a batch size of 256 samples for the RNN model and a token batch size of 4096 for the Transformer. At inference time we ensemble the last three checkpoints to produce the outputs and use beam search with beam size 10.

	de	en	fi	fr	ru	sv
Large	12.0	40.0	7.0	26.0	10.0	3.6
Unidirectional	6.0	20.0	3.5	13.0	5.0	1.8
Edit distance	6.2	18.3	2.6	12.5	3.9	1.1

Table 1: Training set sizes [million sentence pairs].

4. Experiments

Our goal is to generate correct paraphrases for a set of source sentences. However, mere correctness is only part of the story, since generated paraphrases can be very close to the input sentence and as such not that interesting. In the extreme, the output can be identical or almost identical to the input, for instance, *"It is fine."* → *"It's fine."*. Our aim is to produce correct paraphrases, but ideally also paraphrases that are different from their source sentences.
To promote dissimilarity between input source sentences and predicted output sentences, we produce three different versions of the training data:

Large: Our first setup for creating training sets produces the largest number of training examples. For every sentence pair in the training set, we symmetrically train paraphrase generation in both directions, with both sentences as source and target, for instance: *"It stopped raining."* → *"The rain stopped."* and *"The rain stopped."* → *"It stopped raining."*.

Unidirectional: In an attempt to reduce similarity between source and prediction, we use every training sentence pair in one direction only, with random selection of one sentence as the source and the other as the target, for instance: *"The rain stopped."* → *"It stopped raining."*

Edit distance: In a further attempt to promote dissimilarity, we filter the training data to only contain sentence pairs with a minimum edit distance (Levenshtein distance) of 10 or higher.[2] This setup is symmetric, such that both sentences serve as both source and target, for instance: *"I will not let you down."* → *"I won't disappoint you."* and *"I won't disappoint you."* → *"I will not let you down."*.
The training set sizes are shown in Table 1. The unidirectional sets are half has large as the large sets, because every sentence pair occurs in one direction only. The edit distance sets are similar in size to the unidirectional sets.

4.1. Manual annotation

As part of the evaluation, the authors annotated samples of predicted paraphrases. The task was to decide on a four-grade scale how well the predicted sentence was a paraphrase of the input source sentence. The scale consists of the following categories: 1 (Bad), 2 (Mostly bad), 3 (Mostly good), 4 (Good). The same approach was used in the annotation of the Opusparcus development and test sets (Creutz, 2018; Aulamo et al., 2019).

[2]For filtering we use the edit distance figures provided in the Opusparcus data, described in the corpus release: "This adjusted edit distance is computed without taking into account the 'tails' of the longer of the two sentences. For instance, the adjusted edit distance between the sentences 'Frankfurt , Germany .' vs. 'Oh , Frankfurt , Germany .' is zero, because the first shorter sentence fits within the second longer one without any modifications."

	de	en	fi	fr	sv
RNN Large	**98.6**	**100.0**	96.6	**98.6**	**100.0**
RNN Unidirectional	98.3	**100.0**	90.5	93.8	97.9
RNN Edit distance	96.3	98.0	**97.7**	96.1	79.3
Transformer Large	98.5	**100.0**	92.9	91.9	98.4
Transf. Edit Dist.	92.3	96.0	97.6	96.0	74.9

Table 2: Seen test sets: Proportion [%] of predictions annotated as correct ("Good" or "Mostly good" paraphrases). The best result for each language is shown in bold-face font. The methods compared are a traditional encoder-decoder architecture with attention (RNN) on the Large, Unidirectional and Edit distance training sets, and the Transformer on the Large and Edit distance sets.

	de	en	fi	fr	ru	sv
RNN Large	29.3	34.4	21.0	30.8	27.0	18.8
RNN Unidir.	28.7	36.6	21.7	29.0	22.8	14.9
RNN Ed. Dist.	**9.5**	13.7	**5.1**	**7.7**	**4.7**	**1.1**
Transf. Large	30.1	36.7	29.3	28.3	25.8	25.4
Tr. Ed. Dist.	11.0	**11.7**	7.1	8.4	10.8	1.7

Table 3: Seen test sets: Proportion [%] of predictions which are identical copies of the source sentences (the fewer the better).

For each language, two hundred sentence pairs were drawn randomly for annotation, 50 pairs from the *seen test set* and 150 sentence pairs from the *unseen test set* (the latter being more interesting). Manual annotations took place for five languages (all Opusparcus languages except Russian), in five experimental setups: the recurrent architecture with attention trained in turn on the Large, Unidirectional and Edit distance training sets, as well as the Transformer trained on the Large and Edit distance sets.
English and Finnish were annotated independently by two persons. Sentence pairs with too high inter-annotator disagreement were discarded, following the guidelines for the Opusparcus development and test sets (Creutz, 2018; Aulamo et al., 2019). German, French and Swedish were annotated by one person only, so the results for these languages should be seen as indicative at this point.

4.1.1. Seen test sets

Table 2 presents the accuracies of the generated sentences in the seen test sets. Since the input sentences have been seen during training, the accuracies should be high, which is indeed the case in most setups. However, the training sets are noisy. As already mentioned, only 70 % of the sentence pairs that are trained on are estimated to be true paraphrases, so the models are expected to produce errors. In light of this, the accuracies are in fact surprisingly good.
The results in Table 2 suggest that the large (bidirectional) training sets are slightly better than the unidirectional ones. The Transformer does not outperform the RNN on the large sets. The edit distance sets do not reach the same level as the large sets, except for Finnish.
Closer inspection of the results shows that large proportions of the predictions are in fact copies of the source sentence (see Table 3). For English, where we reach 100 % accuracy in theory, more than one third of the predicted outputs are

	de	en	fi	fr	sv
RNN Large	72.4	92.9	**67.4**	**63.9**	**78.1**
RNN Unidir.	70.6	**93.8**	52.5	55.9	66.8
RNN Ed. Dist.	63.6	84.1	55.7	50.4	57.9
Transf. Large	**74.0**	90.8	57.4	60.0	72.2
Transf. Ed. Dist.	68.9	88.8	62.2	54.1	60.5

Table 4: Unseen test sets: Proportion [%] of predictions annotated as correct.

	de	en	fi	fr	ru	sv
RNN Large	2.1	0.9	2.2	3.9	2.8	3.4
RNN Unidir.	2.8	1.0	1.4	3.2	1.8	3.6
RNN Ed. Dist.	**0.0**	**0.2**	**0.8**	**0.2**	**0.6**	0.5
Transf. Large	2.7	2.2	1.9	2.3	0.9	3.0
Tr. Ed. Dist.	0.2	1.1	0.5	0.4	0.9	**0.0**

Table 5: Unseen test sets: Proportion [%] of predictions which are identical copies of the source sentences (the fewer the better).

identical to the inputs. The number of copies is clearly reduced when the edit distance training sets are used, as these models do not encounter output sentences that are very similar to the inputs during training.

4.1.2. Unseen test sets

Tables 4 and 5 show the accuracies of the unseen test sets and the proportions of predictions that are identical copies of the inputs. Compared to the seen test sets, the unseen sentences are naturally more challenging and accuracies are lower. Interestingly, the proportion of copies is also clearly lower. In particular, when trained on the edit distance data, copies almost never occur.

The "RNN Large" setup still produces the highest accuracies on three languages (fi, fr, sv). The Transformer is the best model for German. Unidirectional data seems to work best for English, probably because the English training set is so large that we can afford not duplicating and swapping every sentence pair. Otherwise the unidirectional training approach does not appear too promising, as it mostly does not manage to reduce the proportion of copies compared to the large sets, neither for the unseen (Table 5) nor the seen test data (Table 3). The edit distance models do not reach the accuracy levels of the large models for any language, not even if the copies in Table 5 were to be subtracted from the accuracy figures in Table 4.

Table 6 illustrates another aspect of the paraphrases generated from the unseen test sentences. The figures indicate the proportion of entirely new sentences produced by the models. The values are not particularly high, ranging from 4.8 % (en) to 36.9 % (sv). This means that in the majority of the cases the models are conservative. Rather than inventing a new paraphrase of the source sentence, a target sentence seen during training will be picked. In other words, even if the source sentence is previously unseen, the model will propose a paraphrase from the training set. As the accuracies are rather high (Table 4), this approach often seems to pay off. For instance, the following test source sentences have been paired with a paraphrase observed during training: *"Blew a tire is all."* → *"I got a flat tire."*, *"Things are all changed."* → *"Things are different now."*, and *"I am a citizen of the Federation."* → *"I'm an American citizen."* By contrast, for the following test source sentences new, previously unseen, paraphrases have been created: *"I'll see you, Walter."* → *"See you, Walter."*, *"We thought you ran away from us."* → *"We thought you'd gone."*, *"Point the beam over here."* → *"Give me the beam."*

	de	en	fi	fr	ru	sv
RNN Large	14.7	4.8	19.1	21.0	14.8	16.2
RNN Unidir.	13.3	**7.2**	20.9	19.2	14.6	25.8
RNN Edit Dist.	6.9	4.9	26.5	12.6	20.1	**36.9**
Transf. Large	**16.1**	5.9	13.6	**21.9**	13.6	20.8
Tr. Ed. Dist.	10.6	5.8	**29.5**	14.0	**22.5**	34.2

Table 6: Unseen test sets: Proportion [%] of predictions which are new, i.e. not seen in the training set (the higher, the more "creative" the model is).

4.2. Automatic evaluation

In addition to manual annotation, automatic evaluation has been performed, using three different metrics. Two metrics are designed to measure the semantic adequacy of the output (BLEU and BERTScore), while the third one measures the novelty of the output (PINC). Automatic evaluation of text generation poses a difficult challenge and the choice of an appropriate evaluation metric is not always clear. Consequently, we test how well BLEU and BERTScore correlate with the human annotations described in the previous section to guide future evaluation choices.

BLEU: BLEU is an evaluation metric based on ngram-overlap between the generated candidate sentence and one or more reference sentences (Papineni et al., 2002). It has remained the standard evaluation metric in machine translation. However, in the absence of multiple reference sentences, BLEU can penalize interesting paraphrases that are completely different on the surface despite the same or similar meaning. We report BLEU using both the source and the target sentences from the test data as references. We choose to use the source sentence as reference as well, because similarity to the source can indeed signal semantic adequacy, despite being undesirable in our case.

BERTScore: To remedy the shortcomings of BLEU, we test a recently proposed metric BERTScore that is based on deep contextualized embeddings (Zhang et al., 2020). BERTScore uses a pre-trained BERT model (Devlin et al., 2018) to compute contextualized embeddings for each token. It then calculates pairwise cosine similarities between all candidate and reference tokens, weighted using inverse document frequency. Finally, greedy matching based on the weighted cosine similarities is used for the final score. The process produces three scores: precision, recall, and F1. We use the F1 score in all experiments, as it was found to perform reliably across setups (Zhang et al., 2020). To obtain comparable results across languages, we use the multilingual BERT for our experiments.[3]

[3]We used version 0.1.2 of the BERTScore implementation at `https://github.com/Tiiiger/bert_score` with the

	de	en	fi	fr	ru	sv
RNN Large	34.9	46.6	**25.5**	**33.4**	**33.1**	37.7
RNN Unidir.	32.4	46.4	22.3	30.2	29.2	35.9
RNN Ed. Dist.	24.7	36.0	16.9	18.6	2 2.5	21.7
Transf. Large	**35.9**	**49.0**	22.6	28.3	29.6	**39.4**
Tr. Ed. Dist.	26.5	40.9	18.7	19.3	24.2	22.9

Table 7: BLEU scores on the unseen test sets for all models and languages.

	de	en	fi	fr	ru	sv
RNN Large	**84.8**	84.8	**82.1**	**82.7**	**85.5**	**84.2**
RNN Unidir.	84.3	88.4	80.7	82.2	84.1	83.1
RNN Ed. Dist.	81.8	86.1	78.5	78.4	82.1	79.1
Transf. Large	84.7	**88.6**	81.3	82.1	84.0	83.8
Tr. Ed. Dist.	82.5	86.6	78.9	78.8	82.3	79.4

Table 8: BERTScores on the unseen test sets for all models and languages.

PINC: While BLEU and BERTScore aim to measure the semantic adequacy of the candidate, PINC measures how *dissimilar* the candidate is from the source (Chen and Dolan, 2011). It calculates the percentage of non-overlapping ngrams, essentially being the opposite of BLEU. By combining PINC with the two other metrics we can make sure that models which simply copy the input or perform trivial transformations are penalized.

Automatic evaluation results: Tables 7 and 8 show the BLEU and BERTScores for all models and languages. A look at the tables shows that both metrics rank the models within languages similarly. If we compare the choices of top models based on BLEU and BERTScore, the metrics differ on two languages: In comparison to human judgments, BLEU ranks German correctly and Swedish incorrectly, and BERTScore vice versa. Both metrics rank English incorrectly.

Table 9 shows the PINC scores for all models. As expected, the PINC scores are significantly higher for the edit distance-filtered setups than for the non-filtered ones. A comparison with the BLEU and BERTScores in Tables 7 and 8 shows a trade-off between the semantic adequacy and novelty of the generated paraphrases. For a single evaluation score assessing both the adequacy and novelty, PINC can be combined with one of the other two metrics, for example by calculating an average of the scores. Different weighting schemes for the combination can be used depending on the relative importance of adequacy and novelty for the use case.

Comparison with human judgments: Simply looking at the evaluation scores does not reveal a large discrepancy between BLEU and BERTScore. Therefore we need to analyze more in depth how the metrics compare to human judgments.

First, on corpus-level, that is, the scores in Tables 4, 7 and 8 we find that both metrics correlate well with human evaluation. Using Spearman's rank correlation, we find correlation coefficients of $\rho = 0.898$ for BLEU and $\rho = 0.921$ for BERTScore. Although the difference is not massive, BERTScore clearly outperforms BLEU.

bert-base-multilingual-cased model.

	de	en	fi	fr	ru	sv
RNN Large	71.2	63.1	80.0	66.4	72.8	67.3
RNN Unidir.	70.7	63.7	78.4	66.6	74.4	68.1
RNN Ed. Dist.	**85.1**	**81.1**	**90.8**	84.2	**85.7**	86.0
Transf. Large	68.4	62.4	78.1	67.3	73.6	65.4
Tr. Ed. Dist.	84.6	80.7	88.8	**84.8**	84.6	**88.0**

Table 9: PINC scores on the unseen test sets for all models and languages.

In addition to corpus-level correlation, we test for differences at the sentence level between the two metrics. Figure 1 shows box plots for sentence-level BLEU and BERTScores aggregated from all models for each annotation category. The plots seem to confirm our intuition about the shortcomings of BLEU: While it gives reasonable scores for the lowest category (1: "Bad"), it also gives the full range of scores for paraphrases in the two highest categories (3: "Mostly good" and 4: "Good"). In contrast to BLEU, BERTScore gives few low scores to paraphrases in the higher categories. However, BERTScore generally seems to give higher scores across categories. The lack of lower scores in the lower annotation categories, as well as the similar general trends in scores can also be seen in the plots for Finnish in Figure 2. Overall, the behavior of the two metrics is very different.

In order to quantify the differences between the evaluation metrics on sentence level, we perform two sets of tests. First, we test how well the metrics discriminate between adjacent annotation categories. The Wilcoxon rank-sum test indicates that almost invariably, for both metrics and all adjacent categories, higher scores are given for the higher category ($p < 0.01$). Only two exceptions occur: The BLEU scores for categories 2 and 3 for English and Swedish do not show statistically significant differences at the 0.01 significance level. Based on this, no strong conclusions can be made, although the two non-significant results hint that BERTScore might be a more appropriate sentence-level evaluation metric.

Because the previous test shows no clear difference between the two evaluation metrics, we further test for correlation between the metrics and human judgments using each individual test sentence as a data point. While in the previous test we considered the four annotation categories 1 to 4, we will now include the fine-grained categories 1.5, 2.5 and 3.5 in order to approximate a continuous scale. These categories are the result of two annotators choosing different but adjacent categories for a paraphrase pair. We report results on English and Finnish, since those languages have two annotations for each pair. Using Pearson's *r*, we find that BERTScore correlates better with human judgments (English: $r = 0.44$, Finnish: $r = 0.51$) than BLEU (English: $r = 0.27$, Finnish: $r = 0.40$). Based on these results, we conclude that BERTScore is a more suitable evaluation metric both on corpus level and in the sentence-level scenario.

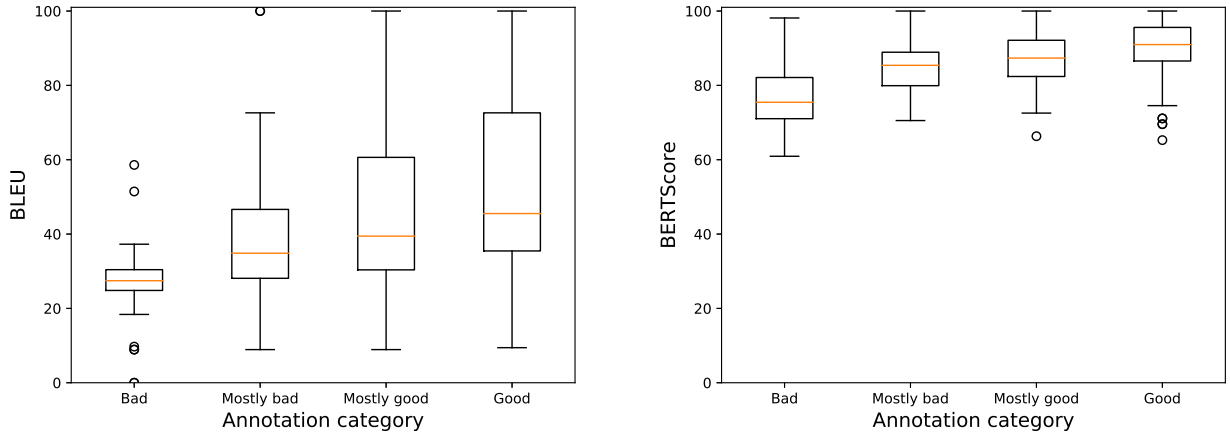

Figure 1: Box plots of BLEU (left) and BERTScore (right) for English showing the distribution of sentence-level scores for each annotation category. The annotation categories are introduced in Section 4.1.

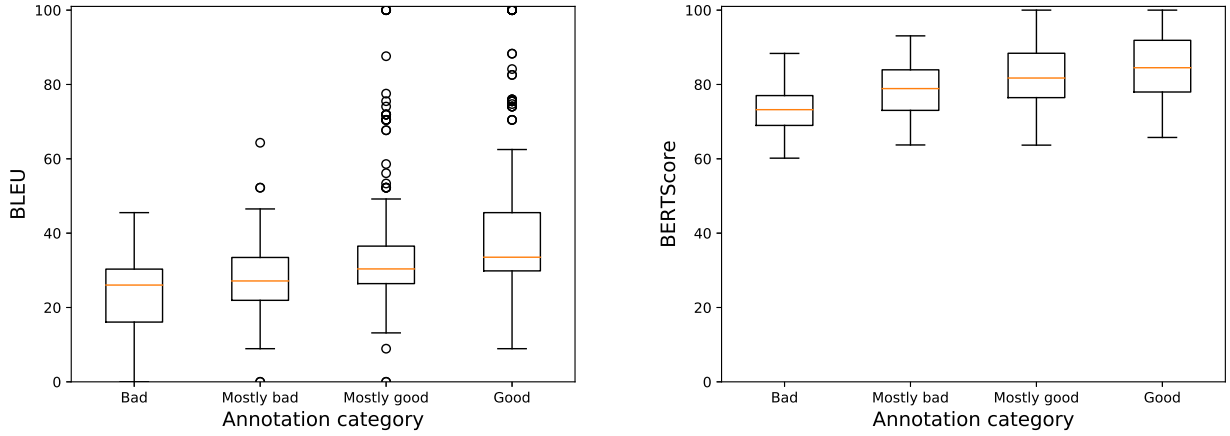

Figure 2: Box plots of BLEU (left) and BERTScore (right) for Finnish showing the distribution of sentence-level scores for each annotation category.

	Source sentence	**Large model prediction**	**Edit distance model prediction**
de	Fahren Sie langsamer. Haben Sie ein gute Nacht.	Fahr langsamer. Gute Nacht.	Nicht so schnell. Schönen Abend noch.
en	Sound familiar? That's the same thing. That was my general impression as well. I did that because it does good to us We thought you ran away from us. So, what brings you to New York?	Does that sound familiar? It's the same thing. That's what I was thinking. It's good for us. We thought you'd gone. What brings you to New York?	Does that ring a bell? Same difference. I thought so too. It's very helpful. I thought you were dead. What brings you to this part of the island?
fi	Haluaisitko jotain? En löydä kenkää.	Haluatko jotain? En löydä niitä.	Voinko tuoda sinulle jotain? Sain potkut.
fr	Depuis quand tu es revenu? Contente de vous revoir.	Depuis quand tu es là? Ravi de vous revoir.	Tu es rentré quand? Bienvenue à la maison.
ru	Мне с тобой скучно. Я деловая женщина.	Мне скучно с тобой. Я же женщина.	Ты утомляешь меня. Я - бизнес.
sv	Nu går jag och lägger mig. Visst gör det ont?	Jag går och lägger mig. Det gör ont, va?	Jag går till sängs. Man blir sårad.

Table 10: Examples of paraphrases generated by the RNN Large and RNN Edit Distance models for all six languages (de, en, fi, fr, ru, sv). In general, the edit distance predictions differ more from the source sentences than the large model predictions do. This creativity can be beneficial, but also sometimes a source of errors or far fetched associations.

5. Discussion and Conclusion

Our experiments have shown that our RNN model (recurrent encoder-decoder with attention trained on bidirectional data) produces the most accurate paraphrase predictions in most cases. We have also shown that it is possible to produce paraphrases that are different from their source sentences, which is a valuable feature for many downstream applications. However, the accuracies obtained for the paraphrases trained on the edit distance training sets are not quite on par with the models trained on the large training sets. Table 10 compares examples of predictions produced by the RNN Large and the RNN Edit Distance models. It appears that the paraphrases produced by the edit distance models can be more interesting, but also semantically more loosely related to the source sentences.

With substantial differences in performance between languages, accuracies range from 64 % to 94 % on unseen data (Table 4). The models turn out to be conservative in the sense that they suggest paraphrases seen during training, whenever appropriate paraphrases can be found in the training set. Completely new sentences are created in roughly 20 % of the cases on average (Table 6).

Our experiments with two metrics for measuring semantic adequacy, BLEU and BERTScore, show good correlations with human assessment especially in corpus-level evaluation. BERTScore, in particular, can be a valuable replacement or complement to labor-intensive manual annotation efforts.

In contrast with recent findings in machine translation, the Transformer-based models did not generally outperform the RNN-based ones in our experiments. This could be partially due to the fact that the RNN models have a slightly higher number of parameters and that the training sets are big enough to learn them reliably. Another possible reason could be that the Transformer output is more creative and would get penalized by BLEU or BERTScore, but this hypothesis is rejected by the PINC results in Table 9.

Comparisons with existing literature on paraphrase generation is difficult because of different data sets and a large variety of evaluation practices. Differences in test sets, for example in terms of the number of available reference sentences, have a large impact on evaluation metrics such as BLEU. While Opusparcus has been used in paraphrase generation before, previous work has used either non-standard train-test splits (Ampomah et al., 2019) or the training sets only (Hämäläinen and Alnajjar, 2019), making their results incomparable to ours. The highest BLEU score reported by Ampomah et al. (2019) for English is 20.1, whereas our corresponding result is 49.0 (Table 7). In a more comparable setting where we calculate BLEU without the source sentence as a reference, our best model still achieves a higher score of 30.3.

Generalization across domains and styles is an open problem, and leveraging multiple data sets from different domains to improve paraphrase generation performance is left for future work. For instance, our best English model trained on Opusparcus scores very low when tested on the Quora Question Pairs corpus (BLEU 10.3) compared to the same model trained on the Quora data (BLEU 34.8). Conversely, a model trained on the Quora data fares even worse on Opusparcus (BLEU 4.84).

In general, we believe that there is still plenty of room for further research on paraphrasing, in particular on natural, colloquial-style data and on languages other than English.

6. Acknowledgments

This study has been supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113). We are grateful to Mikko Aulamo for helping us in the manual annotation effort. We also wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

7. Bibliographical References

Ampomah, I. K., McClean, S., Lin, Z., and Hawe, G. (2019). Jass: Joint attention strategies for paraphrase generation. In *International Conference on Applications of Natural Language to Information Systems*, pages 92–104. Springer.

Aulamo, M., Creutz, M., and Sjöblom, E. (2019). Annotation of subtitle paraphrases using a new web tool. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, Copenhagen, Denmark.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics.

Creutz, M. (2018). Open Subtitles paraphrase corpus for six languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Gupta, A., Agarwal, A., Singh, P., and Rai, P. (2018). A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hämäläinen, M. and Alnajjar, K. (2019). Creative contextual dialog adaptation in an open world rpg. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, United States. ACM.

Hasan, S. A., Liu, B., Liu, J., Qadir, A., Lee, K., Datla, V., Prakash, A., and Farri, O. (2016). Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 42–53.

Iyer, S., Dandekar, N., and Csernai, K. (2017). First Quora dataset release: Question pairs.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 455–462. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Lan, W. and Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Li, Z., Jiang, X., Shang, L., and Li, H. (2017). Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Meteer, M. and Shaked, V. (1988). Strategies for effective paraphrasing. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 431–436. Association for Computational Linguistics.

Nayeem, M. T., Fuad, T. A., and Chali, Y. (2018). Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., and Farri, O. (2016). Neural paraphrase generation with stacked residual LSTM networks. *arXiv preprint arXiv:1610.03098*.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona, Spain. Association for Computational Linguistics.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., and Liu, Y. (2007). Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471.

Roy, A. and Grangier, D. (2019). Unsupervised paraphrasing without translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6033–6039, Florence, Italy. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sjöblom, E., Creutz, M., and Aulamo, M. (2018). Paraphrase detection on noisy subtitles in six languages. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 64–73, Brussels, Belgium. Association for Computational Linguistics.

Suzuki, Y., Kajiwara, T., and Komachi, M. (2017). Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages 36–42.

Vadapalli, R., J Kurisinkel, L., Gupta, M., and Varma, V. (2017). SSAS: Semantic similarity for abstractive summarization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 198–203, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wieting, J. and Gimpel, K. (2017). Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2088, Vancouver, Canada. Association for

Computational Linguistics.

Wieting, J. and Gimpel, K. (2018). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.